RSAC | 2025 Conference

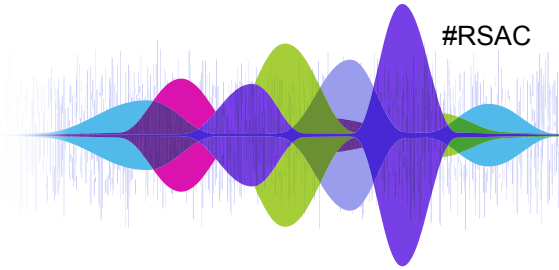
Many Voices.
One Community.

SESSION ID: PDP-R01

Enhancing AI Fairness: Leveraging Differential Privacy

Saahil Shenoy

Founding AI Scientist
Bedrock Security https://www.bedrock.security/
https://www.linkedin.com/in/saahil-shenoy-729a181b/



Disclaimer

Presentations are intended for educational purposes only and do not replace independent professional judgment. Statements of fact and opinions expressed are those of the presenters individually and, unless expressly stated to the contrary, are not the opinion or position of RSA Conference LLC or any other co-sponsors. RSA Conference LLC does not endorse or approve, and assumes no responsibility for, the content, accuracy or completeness of the information presented.

Attendees should note that sessions may be audio- or video-recorded and may be published in various media, including print, audio and video formats without further notice. The presentation template and any media capture are subject to copyright protection.

© 2025 RSA Conference LLC or its affiliates. The RSAC and RSAC CONFERENCE logos and other trademarks are proprietary. All rights reserved.

Two sides of the same coin

- What do the following two have in common?
  - Ensuring customers cannot be re-identified from anonymized dataset
  - Building accurate models for housing prices

Two sides of the same coin

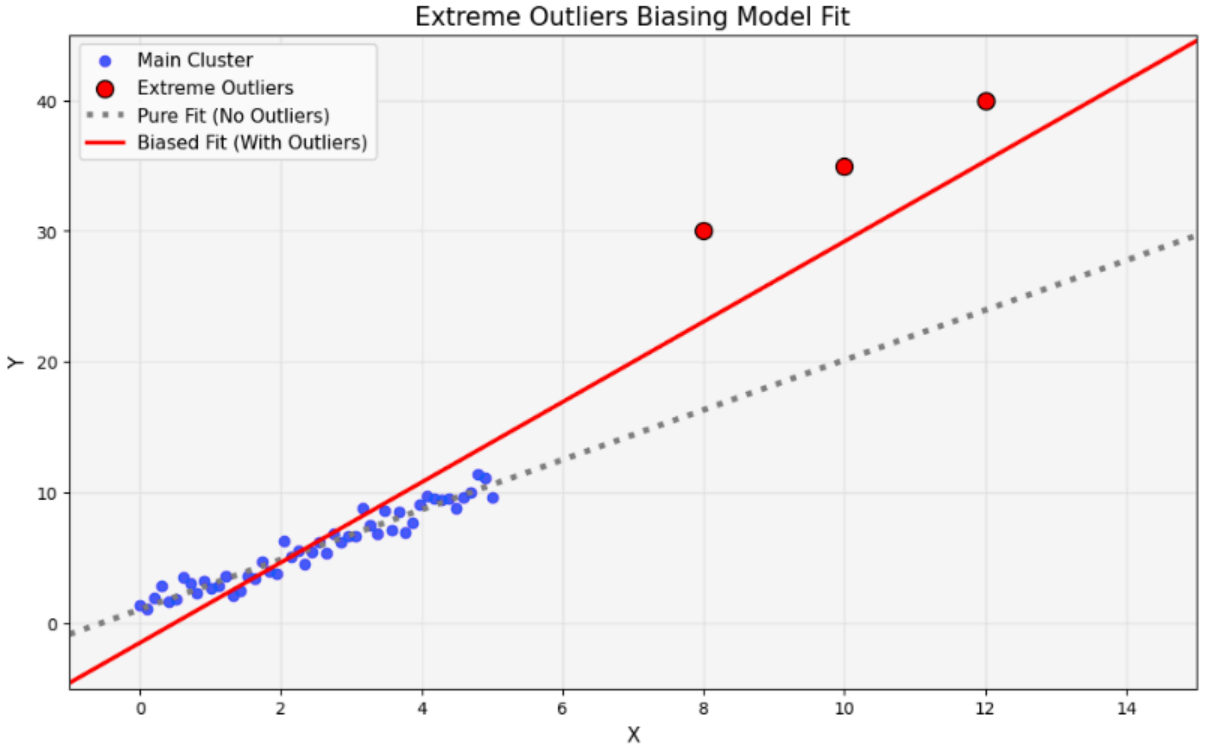
- What do the following two have in common?
  - Ensuring customers cannot be re-identified from anonymized dataset
  - Building accurate models for housing prices
- Hint: It's about AI Fairness

When AI Fairness is not considered

- Incorrect Google Flu Trends (2012-2013)
- Unrealistic Population Projections
- Brittle Financial Models pre-2008

Understanding AI Fairness

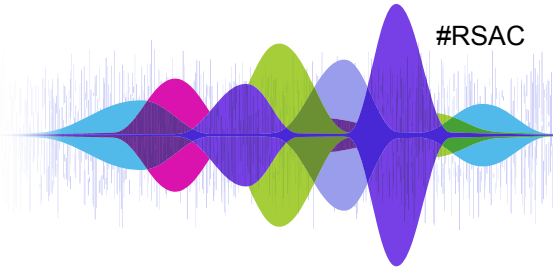
- Outliers can bias time series models
- Predictions can biased on outliers aren't useful for decision making
- Adversarial attacks can manipulate non-robust AI systems more easily



Privacy Concerns

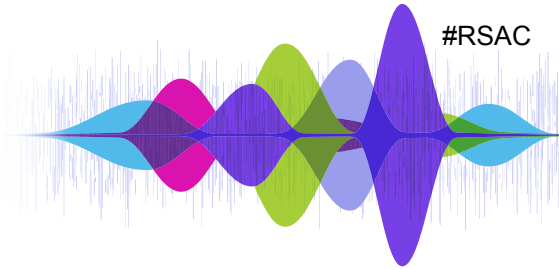
- Most consumers of applications want to be private
- Regulations like GDPR laws push companies to integrate privacy by design
- Reconstruction attacks expose sensitive data increasing need for stronger data security
- Differential privacy (DP) enables ML without compromising privacy

Reconstruction Attacks



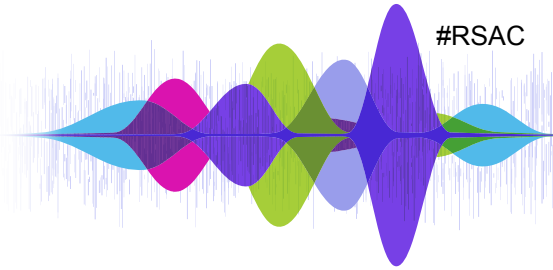
Reconstruction Attacks

- Massachusetts Health Records (1997)
  - Researcher reconstructed Gov. Weld's records using voter data from anonymized medical records



Reconstruction Attacks

- Massachusetts Health Records (1997)
  - Researcher reconstructed Gov. Weld's records using voter data from anonymized medical records
- Netflix Prize Dataset (2006)
  - Anonymized user reviews were re-identified by cross-referencing with IMDb public data



Reconstruction Attacks

- Massachusetts Health Records (1997)
  - Researcher reconstructed Gov. Weld's records using voter data from anonymized medical records
- Netflix Prize Dataset (2006)
  - Anonymized user reviews were re-identified by cross-referencing with IMDb public data
- AOL Search Data Leak
  - NY Times identified a user from their anonymized search queries

RSAC | 2025 Conference

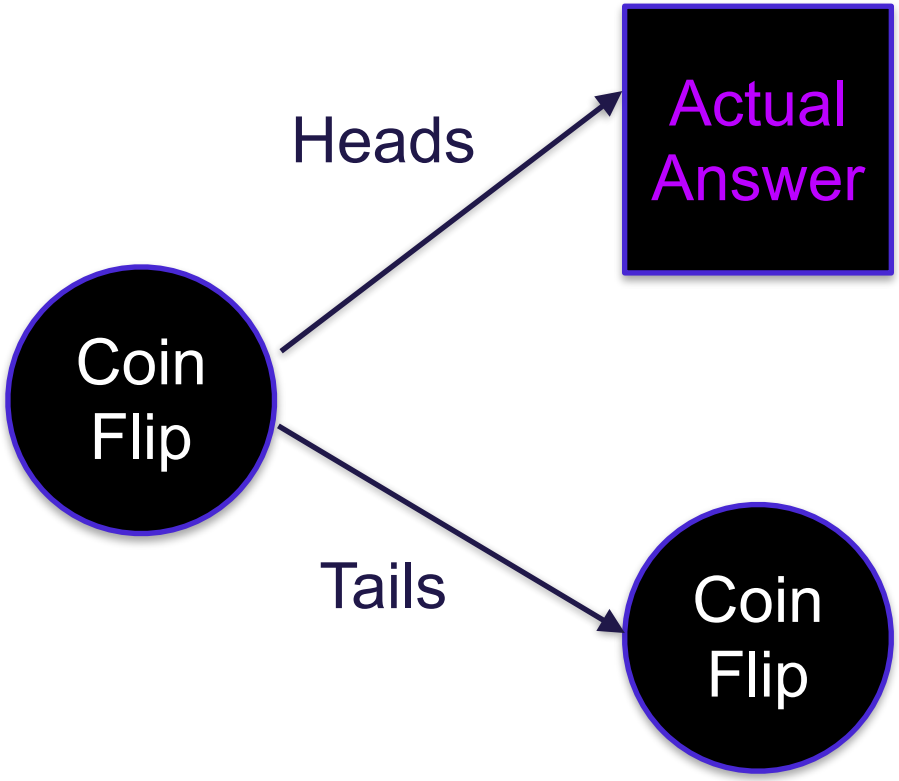
Differential Privacy (DP)

Overview

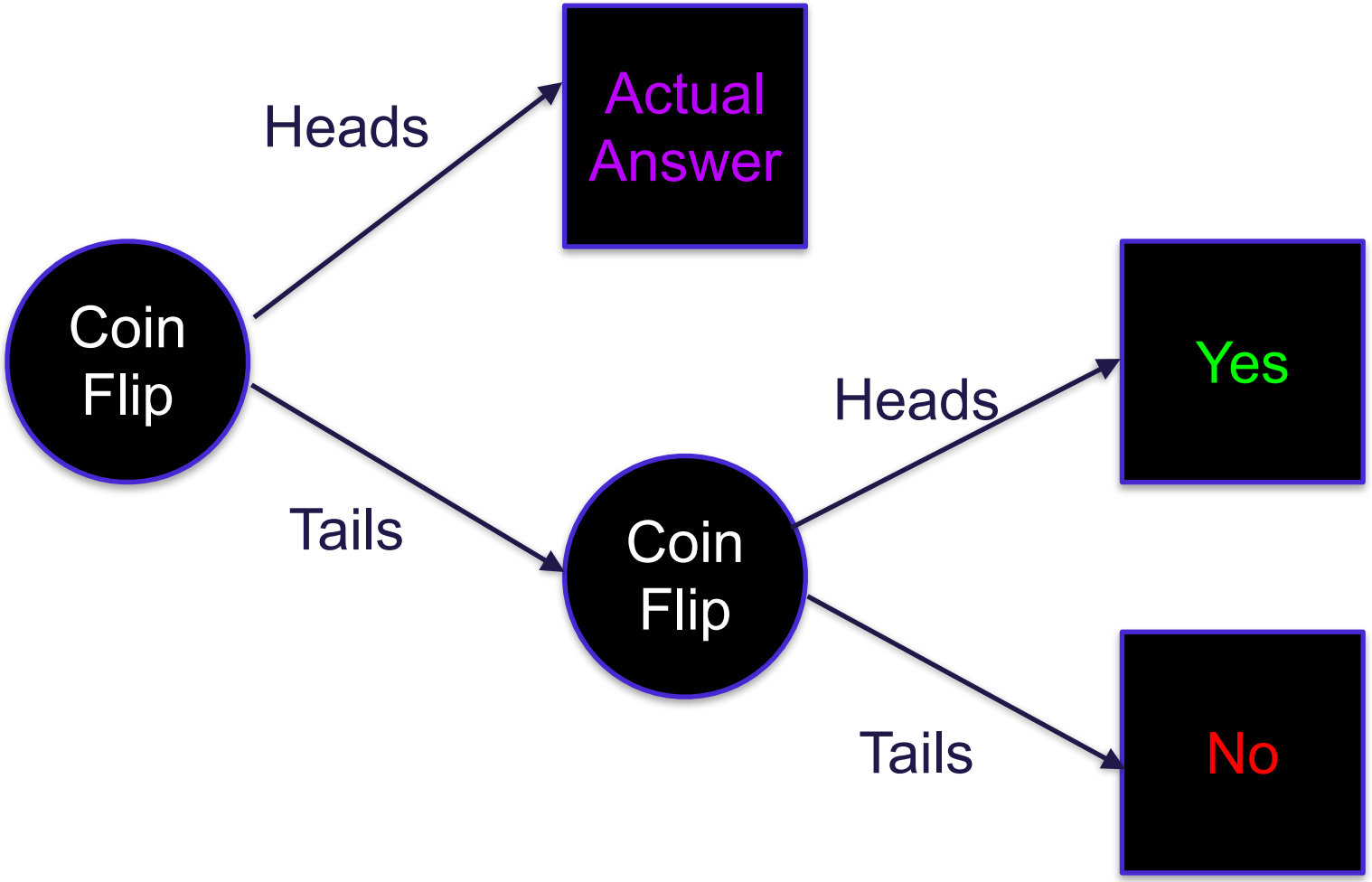
Many Voices.
One Community.

Randomized Response

Randomized Response



Randomized Response



Definition

- Mathematically ensures data insights without revealing any individual's info

Definition

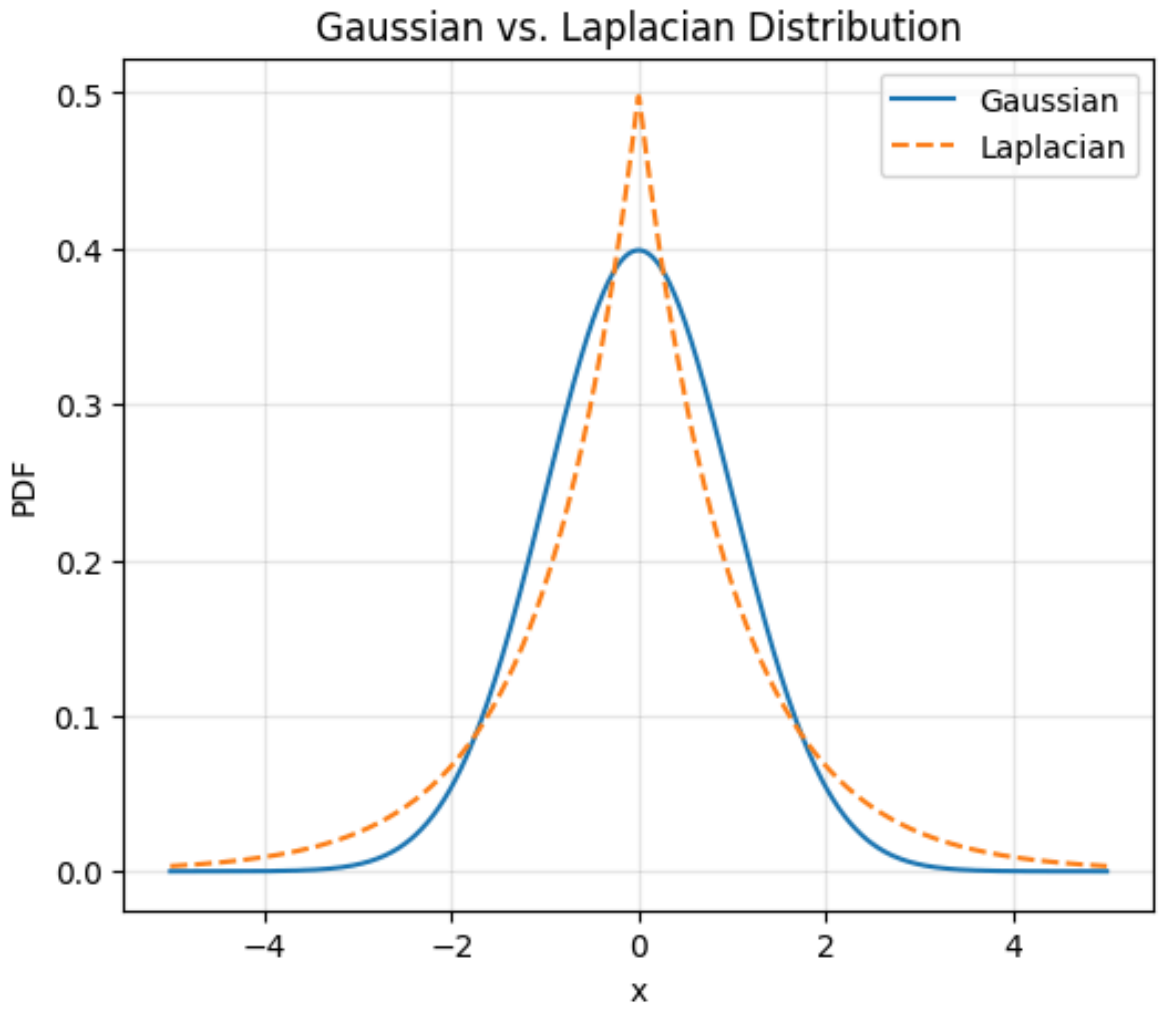
- Mathematically ensures data insights without revealing any individual's info
- Formal definition:

$$\Pr[\mathcal{M}(D_1) \in S] \leq e^{\varepsilon} \Pr[\mathcal{M}(D_2) \in S]$$

  - $\Pr[\ldots]$: Probability function
  - D_1, D_2: Datasets differing by one entry
  - S: Set of possible outputs
  - ϵ: Privacy loss parameter - smaller values means stronger privacy
  - $\mathcal{M}(D_1)$: The algorithm or mechanism that adds noise to the data

Key Mechanisms

- Laplace mechanism
  - Counting, sum, and mean value queries
- Exponential mechanism
  - Selecting most common categorical value
- Gaussian mechanism
  - Relaxed DP



Accuracy-Privacy Tradeoff

- Noise addition
  - Protects privacy but reduces accuracy
- Temporal correlations
  - Needed for accurate forecasting but risks exposing sensitive data
- Privacy-Utility Trade-off
  - Careful tuning required

Global DP

- Query level privacy
- More accuracy, less noise
- Trusted central authority

Local DP

- Individual level privacy
- Less accuracy, more noise
- Higher computation cost

RSAC | 2025 Conference

Applying Differential Privacy

Mitigating Outliers

Many Voices.
One Community.

Differential Privacy on Tabular Data

- Database of individuals' wealth
- Easy to identify individual exists in data when computing mean wealth
- Applying DP mechanism to clip high values and add noise will increase robustness to reconstruction attack

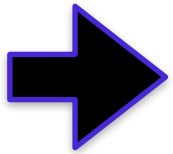
Person	Wealth
Bill Smith	$140k
Billy Bob	$85k
Jeff Billionaire	$1.3B
Brad Ritt	$100k

Differential Privacy on Tabular Data

Person	Wealth
Bill Smith	$140k
Billy Bob	$85k
Jeff Billionaire	$1.3B
Brad Ritt	$100k

Differential Privacy on Tabular Data

Person	Wealth
Bill Smith	$140k
Billy Bob	$85k
Jeff Billionaire	$1.3B
Brad Ritt	$100k



Clip

Person	Wealth
Bill Smith	$140k
Billy Bob	$85k
Jeff Billionaire	$150k
Brad Ritt	$100k

Differential Privacy on Tabular Data

Person	Wealth
Bill Smith	$140k
Billy Bob	$85k
Jeff Billionaire	$1.3B
Brad Ritt	$100k

Clip

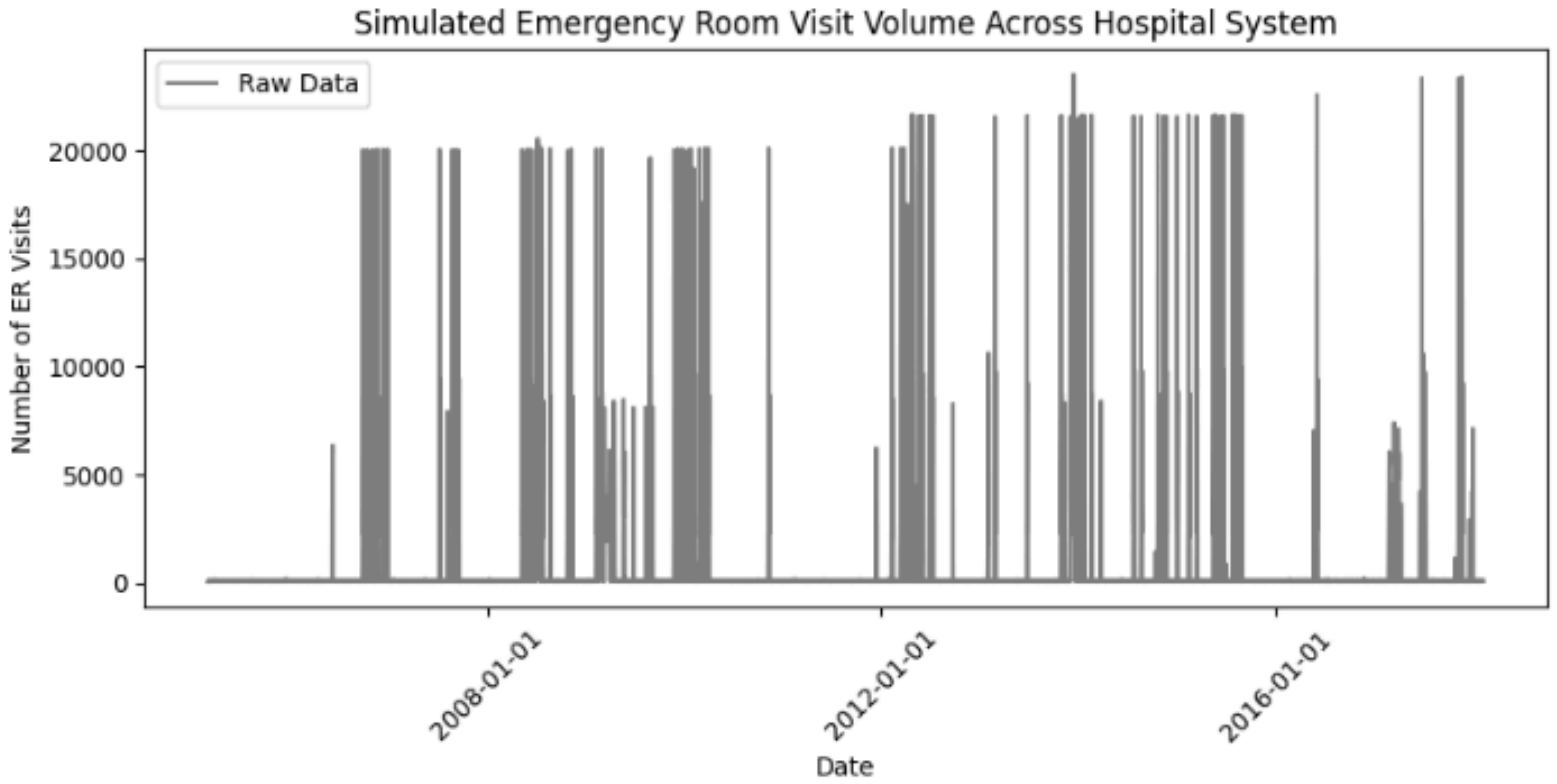
Person	Wealth
Bill Smith	$140k
Billy Bob	$85k
Jeff Billionaire	$150k
Brad Ritt	$100k

Add Noise

Person	Wealth
Bill Smith	$156k
Billy Bob	$75k
Jeff Billionaire	$136k
Brad Ritt	$106k

Case Study: Emergency Room Visits

- Simulated data for number of ER visits in a hospital system
- Data given on a daily basis
- Sparsely populated outliers/spikes in dataset



New Approach Implementation

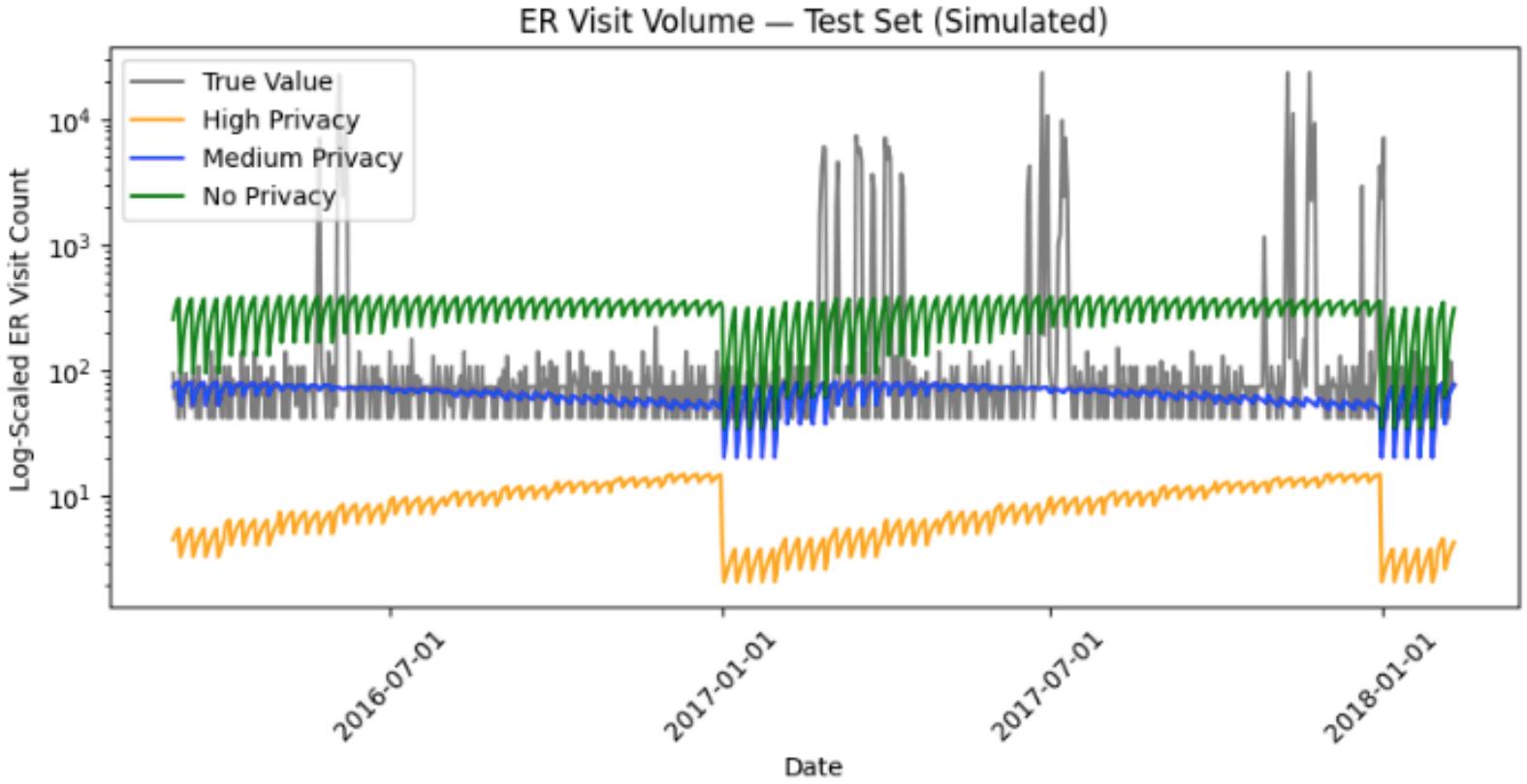
- Combine DP + Recurrent Neural Net (RNN) time series model
- DP stochastic gradient descent (DP-SGD) applied to the modeling training
  - Involves limiting (clipping) magnitude gradient steps
  - Can only use clipped values of time series for training
- Only have access to timestamp information during inference

Results

Parameter Values

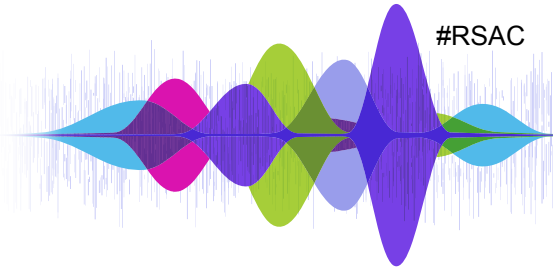
	ε	δ
High	0.5	1E-05
Med	3	1E-05
None	∞	1

- ϵ : controls how much noise is added - smaller means more noise
- δ : Very small probability that privacy guarantee might not hold



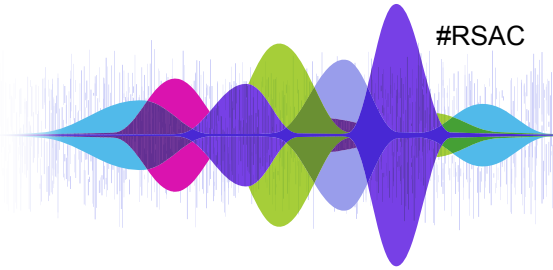
RSAC | 2025 Conference

How to Apply Differential Privacy for Your Organization



Open-Source DP Tools

- **Google's DP Libraries**: offered in C++, Go, and Java
- **OpenDP**: collaboration between Harvard and Microsoft
- **IBM's Diffprivlib**: integrates with NumPy & scikit-learn
- **Tensorflow Privacy**: developed by Google
- **Opacus**: developed by Meta and integrates with PyTorch
- **PySyft**: designed by OpenMined integrates w/PyTorch & Tensorflow



Integrating DP into ML Workflows

- Use DP-SGD for Training: Clip gradients, add noise
- Track Privacy Budget: Balance privacy vs. accuracy
- Leverage DP Libraries: TensorFlow Privacy, Opacus, PySyft, etc.
- Simplify & Tune Hyperparameters: Reduce sensitivity, optimize settings
- Report & Document: Transparency in DP choices

Apply What You Have Learned Today

- Next week you should:
  - Identify your organization's most sensitive and critical datasets used in your AI models
- In the first three months following this presentation you should:
  - Map who accesses these datasets and why
  - Assess privacy risks and where DP can help
- Within six months you should:
  - Select DP framework or tool
  - Launch initiative to integrate privacy-preserving controls into AI workflows

Looking Ahead

- Research opportunities
  - Exploring advanced noise mechanisms or adaptive privacy budgets
- Policy Implications
  - Discussing the need for the guidelines on privacy-preserving data analysis
- Adoption Strategy
  - To get started, start with data discovery and roll out DP to where sensitive data is being used most

RSAC | 2025 Conference

Many Voices.
One Community.

Thank You

For more information visit

https://run.bedrock.security/differential-privacy-and-ai-at-rsa-2025?hs_preview=pvlPXxEE-173436496799